False Dichotomy Alert: Improving Subjective-Probability Estimates vs. Raising Awareness of Systemic Risk

Philip E. Tetlock[a, **] (tetlock@wharton.upenn.edu), Yunzi Lu[a] (yunzilu@sas.upenn.edu), Barbara A. Mellers[a] (mellers@wharton.upenn.edu)

[a] Wharton School, University of Pennsylvania, Philadelphia PA 19104, USA

** Corresponding author. E-mail address: tetlock@wharton.upenn.edu (Philip E. Tetlock)

Abstract

Taleb et al. (2022) portray the superforecasting research program as a masquerade that purports to build "survival functions for tail assessments via sports-like tournaments.'' But that never was the goal. The program was designed to help intelligence analysts make better probability judgments, which required posing rapidly resolvable questions. From a signal detection theory perspective, the superforecasting and Taleb et al. programs are complementary, not contradictory (a point Taleb and Tetlock (2013) recognized). The superforecasting program aims at achieving high Hit rates at low cost in False-Positives whereas Taleb et al prioritize alerting us to systemic risk, even if that entails a high False-Positive rate. Proponents of each program should however acknowledge weaknesses in their cases. It is unclear: (a) how Taleb et al. (2022) can justify extreme error-avoidance trade-offs, without tacit probability judgments of rare, high-impact events; (b) how much superforecasting interventions can improve probability judgments of such events.

Keywords: superforecasting, systemic risk, fat-tailed dstributions, signal detection, forecasting tournaments, proper scoring rules

False Dichotomy Alert: Cultivating Talent at Probability Estimation vs. Raising Awareness of Systemic Risk

Taleb et al. (2022, p.3) call the "superforecasting" research program a "masquerade"—a term that the Oxford English Language Dictionary defines as "a false show or pretense" and a characterization that some readers might interpret, not unreasonably, as an accusation of fraud. The "superforecasters masquerade" allegedly involves "building survival functions for tail assessments via sports-like ''tournaments,'' instead of "using more rigorous approaches like Extreme Value Theory." The tournament approach, which they attribute to Tetlock and Gardner (2016) is, "simply wrong and violates elementary probability theory."[1]

We divide our response into two parts. First, we clarify the long-standing goals of the superforecasting research program. Second, we use a signal detection framework to explain why the superforecasting program and Taleb et al. program are complementary, not contradictory. Recognizing this complementarity will however require each side to recognize weaknesses in its case. There are sound statistical reasons for expecting the power of superforecaster-style interventions to improve probability judgments will fall toward zero for increasingly rare events. And there are compelling logical grounds for supposing that Precautionary-Principle recommendations on where to set systemic-risk alarm thresholds ultimately rest on tacit probability judgments.

I. Clarifying Goals of Superforecasting Research Program. The Taleb et al. critique rests on a misconception. The superforecasting research program was not designed to build survival functions for tail assessments. It emerged from forecasting tournaments that the U.S. intelligence

[1] Curiously, Taleb (2007, page 210-211), in The Black Swan, praised Tetlock's (2005) earlier forecasting tournaments (using Brier scores) for providing evidence that political experts are poorly calibrated, which raises questions about shifting standards of proof: Why can't later forecasting tournaments (also using Brier scores) generate equally convincing evidence that talented generalists can be well-calibrated in their confidence judgments?

community sponsored with a very different goal in mind: improving intelligence analysts' probabilistic answers to geopolitical questions that could be rigorously resolvable in the near future, thus enabling rapid accuracy feedback. The superforecasting program won these tournaments against serious competitors by substantial margins—and it did so by doing a better job at spotting talent, cultivating talent and aggregating crowd forecasts (Tetlock & Gardner, 2016). If this were a masquerade, the Office of the Director of National Intelligence and the Intelligence Advanced Projects Activity (IARPA) would have had to be colluding in the deception. IARPA is committed to Open Science norms and it ensured impartiality in the collection of data and scoring of forecasts as well as in the archiving of data for public inspection. We challenge Taleb et al. (2022) to be equally transparent about the performance of their tail-risk hedging strategies—a controversial topic (Brown, 2020).

To achieve its essentially psychometric objectives, the IARPA program picked forecasting questions in the Goldilocks zone of difficulty, neither so hard that virtually no one could get better (e.g., will the S&P 500 index rise or fall tomorrow?) nor so easy that virtually everyone would get everything right (e.g., will more than 100 million perish in a nuclear war tomorrow?). To this end, they posed questions that expert question generators classified, ex ante, as falling in the 10% to 90% range of likelihood of occurrence: estimates of Assad regime collapse in 2013 or of German-Spanish bond-yield spreads during the Grexit crisis or of casualty counts linked to naval confrontations in the South or East China Seas in 2011.

It was on these types of questions that the superforecasting research program chalked up its successes: identifying systematic individual differences in forecasting strategies and performance (Atanasov et al., 2020; Bo et al., 2017; Chen et al., 2016; Friedman et al., 2018; Horowitz et al., 2019; Karvetski et al., 2021; Mellers et al., 2014; Mellers, Stone, Atanasov, et al., 2015; Mellers, Stone, Murray, et al., 2015; Mellers et al., 2017; Mellers et al., 2018; Merkle et al., 2017; Moore et

al., 2017; Schwartz et al., 2017; Tetlock & Gardner, 2016), developing better algorithms for distilling wisdom from crowds (Baron et al., 2014; Cross et al., 2018; Satopää et al., 2014; Satopää et al., 2021), fine-tuning methods of training novice forecasters (Chang et al., 2016, Chang et al., 2017) and of incentivizing constructive contributions to debate (Chang et al., 2017; Tetlock et al., 2014) and assessing the relative merits of eliciting forecasts via self-report tournaments versus pricing judgments in prediction markets (Atanasov et al., 2017; Dana et al., 2019; Inchauspe et al., 2014; Mellers & Tetlock, 2019). The average Brier scores on 400-plus questions across four years ranged from 25% to 70% lower than those for the official IARPA-designated control group, for our scientific competitors and for a prediction market using intelligence analysts with access to classified information (Goldstein et al., 2016).

There have now been over well over 20 publications in peer-reviewed journals involving more than 30 collaborators. If superforecasting is a "masquerade," it implicates not only the U.S. intelligence community but at least 10 peer-reviewed journals (including the International Journal of Forecasting) and investigators across several major universities. If Taleb et al. are right in their sweeping dismissal of the superforecasting program, they have identified a massive lapse of judgment of funding organizations, journals and universities. Fortunately, the lapse of judgment is more localized. Indeed, we can pinpoint where Taleb et al. go wrong.

In adopting so dismissive a stance, Taleb et al. put, in effect, zero value on providing policy-makers with more accurate answers to IARPA-style questions. We understand the temptation. Predicting small-scale naval clashes in the South China Sea looks trivial in relation to reducing the systemic risk of, say, a Sino-American war, which could kill 600 million, at least 10X the death toll in World War II. This argument assumes, however, a false dichotomization of research options. We do not have to choose between improving short-run probability judgments and contingency planning for systemic risks. Taleb et al.'s (2022) writing off the superforecasting

program is analogous to writing off the work of vision scientists who invent methods of improving human eyesight by 30% but—cue complaint—we still cannot spot predators lurking miles away. If Taleb et al. (2022) cannot conceive of how boosting accuracy for near-run events might be useful to policy-makers—say, by highlighting early-warning indicators of catastrophe—that is a failure of imagination on their part, not a failure of the superforecasting program.

II. Complementary not Contradictory Goals. Taleb et al. (2022, p.3) do concede that "just as there are frivolous lawsuits, there are frivolous risk claims and… we limit these precautionary considerations to a precise class of fat-tailed multiplicative processes – when there is systemic risk." This concession is consequential: it opens the door to the sort of conceptual integration that Taleb and Tetlock (2013) attempted but failed to achieve.

Signal detection theory (SDT) is a useful starting point for this integration. In SDT, Receiver Operating Characteristic (ROC) curves measure a forecaster's skill at distinguishing false-positive from true-positive signals (Green & Swets, 1966). Let "frivolous lawsuits" be the false-positive signals, analogous to "crying wolf" when there is no wolf. And let prudent precautionary responses to real threats be the true-positive signals. The ROC curves in Figure 1 display the false-positive vs. true-positive trade-offs possible for forecasters with known levels of predictive accuracy in an IARPA-style tournament, where event base rates cluster in the 10-90% range. The faster a curve rises above the chance-performance diagonal, the more accurate the forecasters. The blue superforecasting curve, with a d-prime of 1.0, has more AUC than the gray regular-forecaster curve, with a d-prime of 0.5 (both values are rough approximations of reality—Mellers et al. 2015). Points along a curve represent the set of possible tradeoffs between false-positives and true-positives for a given level of predictive accuracy. Consider the blue points A, B and C on the superforecasting ROC curve—and the grey points A', B', C' on the regular-forecaster curve. At point A on the superforecasting curve, the price of a 50% true-positive rate is a 16% false-positive

rate; at point B, the price of an 84% true-positive rate is a 50% false-positive rate; at point C, the price of a 99% true-positive rate is a 90% false-positive rate. We have to pay steeper prices on the regular-forecaster curve. At A', the price of a 50% true-positive rate rises to a 31% false-positive rate; at B', the price of an 84% true-positive rate rises to a 67% false-positive rate; at C', the price of a 99% true-positive rate rises to a 96% false-positive rate. The mission of the superforecasting program is to raise ROC curves and increase AUC. Superforecasting is, in this sense, value-neutral. It does not presume to tell decision makers how to trade off true- against false-positives—and when to ring the systemic-risk alarm bell.

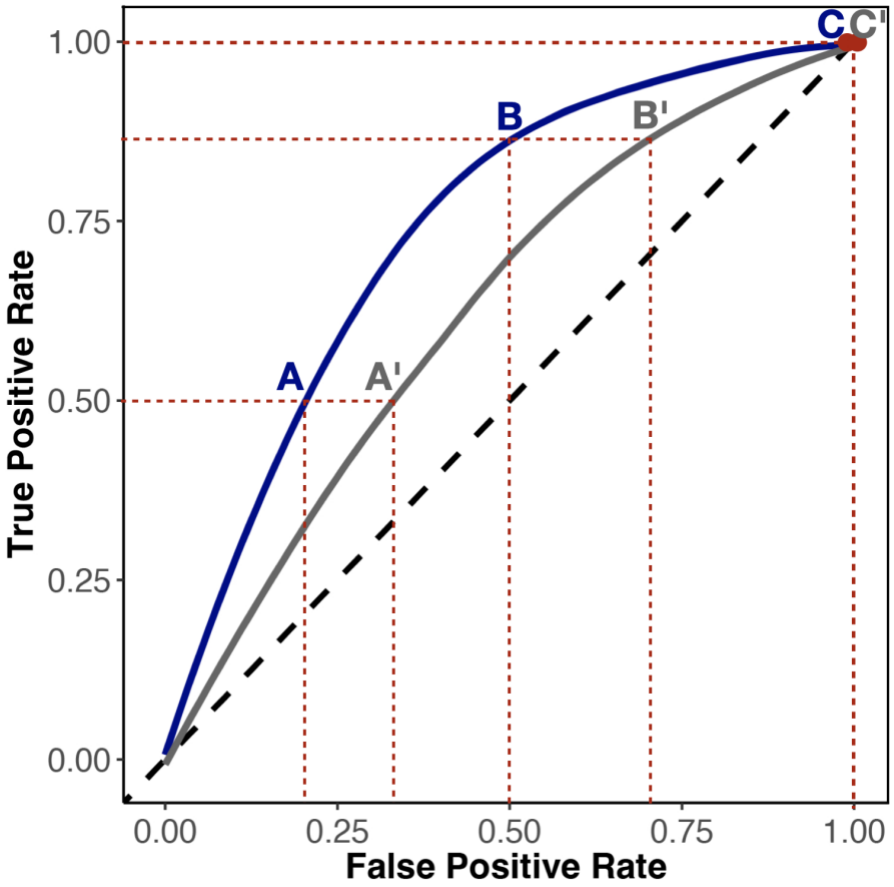

Figure 1. Hypothetical ROC curves, based on Mellers et al. (2015), capture complementarity of research programs aimed at raising accuracy and at managing extreme risks: the more area under the curve (AUC), the more successful the raising-accuracy program; the wider the threshold shift in tolerance of false-positives, the more impactful the Talebian program.

Although there are opposing views on the value of IARPA-style tournaments, there is substantial scientific consensus on the value of SDT. It defines sharp fact-value divisions of labor: in this case, between analysts whose job is making accurate forecasts and policymakers whose job is making value judgments (setting thresholds for issuing alarms). We see SDT as a cornerstone on which clashing schools of thought could build but only if each is willing to take the first step toward successful adversarial collaborations (Mellers et al., 2001; Kahneman & Klein, 2009). That step is to convince the other side that one is arguing in good faith—and understands its core concerns.

It is in this spirit we try here to characterize (not caricature) the core objections that Taleb et al. (2022) have to our program. We see two. First, they doubt the value of improving signal detection outcomes in short-range IARPA tournaments. A 10% accuracy boost in forecasts of incremental variation in global-surface temperature next year matters naught if superforecasters miss the outlier .0001 chance of a sudden 3-degree C spike. Second, they understandably worry that as soon as superforecasters leave Mediocristan (Taleb, 2007), the idealized Gaussian realm of SDT, and enter Extremistan, the domain of fat tails, they will be in for a rude shock. The classic SDT accuracy metric, AUC, becomes unstable—and eventually meaningless.[2]

In the same spirit, we lay out our two core concerns, without resort to caricature. First, Taleb et al. offer too sweeping a critique of meliorist efforts to improve probability judgments—and seem to make no allowance for empirical surprises. But there is much we do not know. No one has ever tested the capacity of the superforecasting selection-and-training programs to produce teams

[2] Ferro and Stephenson (2011) propose two indices to compensate for the shortcoming of SDT metrics in tail-risk domains, the extremal dependence index and the symmetric extremal dependence index, which they argue are non-degenerating, base-rate independent, asymptotically equitable, and harder to hedge.

that can adapt to tail-risk challenges.[3] Although our critics could be right that meliorist research programs for boosting accuracy will stall out for increasingly rare events, neither we nor they know where the point of diminishing marginal predictive returns lies. We also wonder whether our critics are making insufficient allowance for the capacity of statisticians to develop metrics, like the symmetric extremal dependence index (Ferro & Stephenson, 2011), that are more robust to sparse data than are traditional SDT metrics.

Second, by categorically rejecting meliorist efforts to boost accuracy, Taleb et al. risk undercutting Precautionary-Principle positions in real-world policy debates that prominent Extreme Value theorists have endorsed. Categorical rejection implies that it is impossible to make the probabilistic distinctions essential for winnowing out false-positives ("frivolous lawsuits") and leaves us with little recourse but to adopt highly skewed error-avoidance functions. Although this better-safe-than-sorry position may look prudent, it does not hold up under close inspection. Without a winnowing mechanism, we grant de facto veto power to proponents of the strong Precautionary Principle over any technology whenever a vocal constituency can conjure up a worst-case story about systemic risk—from genetically modified crops to the Large Hadron Collider. The question becomes: how implausible/improbable must such stories become to make even Extreme Value theorists balk? Formal mathematical arguments about "fat-tailed multiplicative processes" do not answer this question—and should not obscure how often reasonable people disagree on how to answer it.

Closing Remarks. We welcome constructive critiques grounded in an understanding of the objectives of the superforecasting program. Value-neutral Brier scoring is one of a large class of

[3] Karger, Atanasov and Tetlock (2022) are launching a series of studies to test the skills of both superforecasters and subject-matter experts at assigning realistic probabilities to a wide range of short- and longer-term X-risk indicators. In the spirit of Kahneman-style adversarial collaborations, we welcome the participation of Extreme Value theorists in these exercises—as either forecasters or as acceptable-risk threshold setters.

reasonable metrics for gauging improvements in probability estimates of geopolitical events in the near future—e.g., low-casualty skirmishes in South China Sea—that are potential early warning indicators of systemic risk. No one disputes the need for different metrics if our goal becomes constructing survival functions for tail-risk assessments of, say, a Sino-American war that claims millions of lives: a shift toward eliciting continuous probability distributions of outcomes and adopting proper scoring rules tailored to decision-makers' utility functions, such as logarithmic scoring that is extremely punitive toward missing tiny-probability events (Karger et al., 2022). Different research goals call for different research methods. It would be silly for us to criticize the Taleb et al.'s (2022) program for not offering guidance on how to win an IARPA-style tournament. And it is equally misguided for Extreme Value theorists to criticize us for not telling policy-makers how to hedge their exposure to tail risks.

Future generations of forecasting scholars will succeed where Taleb and Tetlock (2013) failed—and bridge this false dichotomy. Efforts to improve early-warning systems can co-exist comfortably with sensitivity to systemic risk. As we should have learned from the COVID-19 pandemic—a catastrophe that micro-biology textbooks warned of 20 years ago—it is useful to array systemic risks along a White-to-Gray-to-Black-Swan continuum (Tetlock, 2017). As swans go, the COVID-19 pandemic was only grayish, closer to White than Black. It offered up many warning signals that spiked in intensity in the fall of 2019 and that we have paid a steep price for ignoring. We agree with Taleb et al. (2020) that society should be doing much more contingency planning to pre-empt systemic risks. But contingency planning that goes beyond rhetorical posturing becomes very expensive very fast. We cannot go full throttle on pre-empting every candidate threat that pops up on someone's radar screen. We must set priorities. And those priorities must rest, in part, on implicit probability judgments of expected impacts, which means there is potential value in improving the accuracy of those judgments in the very forecasting

exercises that Taleb et al. (2022) disparage. Consider: what would a prudent President of the United States have done differently if he had adjusted his probability of a 2020 pandemic from 1% to 10% in response to CIA briefings on early warning indicators in the fall of 2019? And then ask: would a prudent President have been likelier to act on the briefings if he had had quantitative evidence on the predictive track records of the briefers? The answers underscore the urgency of institutionalizing forecasting tournaments across a wide range of organizations—and of systematically linking early warning indicators to longer-run systemic risks.[4]

[4] The failure of the Taleb and Tetlock (2013) collaboration also has philosophical roots. Anti-meliorists insist, with mathematical certitude, that the improving-accuracy agenda is doomed as soon as it moves from forecasting short-run outcomes in the 10%-to-90% zone to outcomes in the tails of probability distributions. We may have to wait centuries to discover which forecasters are well calibrated—and even those results will be limited to a quirky single run of history. No one will ever know how "lucky" we have been to have avoided World War III, so far. The meliorist agenda is thus a net-harmful distraction from the greater-good goal of pre-empting compelling long-run systemic risks. However, meliorists counter that judgments of "compellingness" rest on covert probability judgments and pretending otherwise makes it too easy for demagogues to dominate systemic-risk debates. If we don't want debates to reduce to leaps of faith, we should try to improve the tail-risk judgments informing policy priorities. For instance, Karger et al. (2022) offer an unapologetic defense of meliorism that identifies promising research strategies for improving judgments of existential risks.

Acknowledgements: We acknowledge the very helpful comments of Pavel Atanasov as well as the research support of three foundations: Founders Pledge, Open Philanthropy and Templeton. No Conflicts of Interest.

References

Atanasov, P., Rescober, P., Stone, E., Swift, S. A., Servan-Schreiber, E., Tetlock, P., Ungar, L., & Mellers, B. (2017). Distilling the wisdom of crowds: Prediction markets vs. prediction polls. *Management Science, 63*(3), 691-706. https://doi.org/10.1287/MNSC.2015.2374

Atanasov, P., Witkowski, J., Ungar, L., Mellers, B., & Tetlock, P. (2020). Small steps to accuracy: Incremental belief updaters are better forecasters. *Organizational Behavior and Human Decision Processes, 160*, 19-35. https://doi.org/10.1016/J.OBHDP.2020.02.001

Baron, J., Mellers, B. A., Tetlock, P. E., Stone, E., & Ungar, L. H. (2014). Two reasons to make aggregated probability forecasts more extreme. *Decision Analysis, 11*(2), 133-145. https://doi.org/10.1287/DECA.2014.0293

Bo, Y. E., Budescu, D. V., Lewis, C., Tetlock, P. E., & Mellers, B. (2017). An IRT forecasting model: Linking proper scoring rules to item response theory. *Judgment and Decision Making, 12*(2), 90-103.

Brown, A. (2020). *Taleb-asness black swan spat is a teaching moment*. Bloomberg. https://www.bloombergquint.com/gadfly/twitter-spat-over-market-risks-is-a-teaching-moment

Chang, W., Atanasov, P., Patil, S., Mellers, B. A., & Tetlock, P. E. (2017). Accountability and adaptive performance under uncertainty: A long-term view. *Judgment and Decision Making, 12*(6), 610-626.

Chang, W., Chen, E., Mellers, B., & Tetlock, P. (2016). Developing expert political judgment: The impact of training and practice on judgmental accuracy in geopolitical forecasting tournaments. *Judgment and Decision Making, 11*(5), 509-526.

Chen, E., Budescu, D. V., Lakshmikanth, S. K., Mellers, B. A., & Tetlock, P. E. (2016). Validating the contribution-weighted model: Robustness and cost-benefit analyses. *Decision Analysis, 13*(2), 128-152. https://doi.org/10.1287/DECA.2016.0329

Cross, D., Ramos, J., Mellers, B., Tetlock, P. E., & Scott, D. W. (2018). Robust forecast aggregation: Fourier L2E regression. *Journal of Forecasting, 37*(3), 259-268. https://doi.org/10.1002/FOR.2489

Dana, J., Atanasov, P., Tetlock, P., & Mellers, B. (2019). Are markets more accurate than polls? The surprising informational value of "just asking". *Judgment and Decision Making, 14*(2), 135-147.

Ferro, C. A., & Stephenson, D. B. (2011). Extremal dependence indices: Improved verification measures for deterministic forecasts of rare binary events. *Weather and Forecasting, 26*(5), 699-713. https://doi.org/10.1175/WAF-D-10-05030.1

Friedman, J. A., Baker, J. D., Mellers, B. A., Tetlock, P. E., & Zeckhauser, R. (2018). The value of precision in probability assessment: Evidence from a large-scale geopolitical forecasting tournament. *International Studies Quarterly, 62*(2), 410-422. https://doi.org/10.1093/ISQ/SQX078

Goldstein, S., Hartman, R., Comstock, E., & Baumgarten, T. S. (2016). *Assessing the accuracy of geopolitical forecasts from the US Intelligence Community's prediction market.* Working Paper.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. Wiley.

Horowitz, M., Stewart, B. M., Tingley, D., Bishop, M., Samotin, L. R., Roberts, M., Chang, W., Mellers, B., & Tetlock, P. (2019). What makes foreign policy teams tick: Explaining variation in group performance at geopolitical forecasting. *Journal of Politics, 81*(4), 1388-1404. https://doi.org/10.1086/704437

Inchauspe, J., Atanasov, P., Mellers, B., Tetlock, P., & Ungar, L. (2014). A behaviorally informed survey-powered market agent. *The Journal of Prediction Markets, 8*(2), 1-28. https://doi.org/10.5750/JPM.V8I2.867

Karger, E., Atanasov, P., & Tetlock, P. E. (2022). Improving judgments of existential risk. *SSRN Working Paper*, Article 4001628. http://dx.doi.org/10.2139/ssrn.4001628

Karvetski, C. W., Meinel, C., Maxwell, D. T., Lu, Y., Mellers, B. A., & Tetlock, P. E. (2021). What do forecasting rationales reveal about thinking patterns of top geopolitical forecasters? *International Journal of Forecasting*, *38*(2), 688-704. https://doi.org/10.1016/J.IJFORECAST.2021.09.003

Kahneman, D., & Klein, G. (2009). Conditions for intuitive expertise: A failure to disagree. *American Psychologist, 64*(6), 515–526. https://doi.org/10.1037/a0016755

Mellers, B.A., Hertwig, R., & Kahneman, D. (2001). Do frequency representations eliminate conjunction effects? An exercise in adversarial collaboration. *Psychological Science, 12*, 269-275.

Mellers, B., Stone, E., Atanasov, P., Rohrbaugh, N., Emlen Metz, S., Ungar, L., Bishop, M. M., Horowitz, M., Merkle, E., & Tetlock, P. (2015). The psychology of intelligence analysis: Drivers of prediction accuracy in world politics. *Journal of Experimental Psychology: Applied, 21*(1), 1-14. https://doi.org/10.1037/XAP0000040

Mellers, B., Stone, E., Murray, T., Minster, A., Rohrbaugh, N., Bishop, M., Chen, E., Baker, J., Hou, Y., Horowitz, M., Ungar, L., & Tetlock, P. (2015). Identifying and cultivating superforecasters as a method of improving probabilistic predictions. *Perspectives on Psychological Science, 10*(3), 267-281. https://doi.org/10.1177/1745691615577794

Mellers, B. A., & Tetlock, P. E. (2019). From discipline-centered rivalries to solution-centered science: Producing better probability estimates for policy makers. *American Psychologist, 74*(3), 290-300.

Mellers, B., Ungar, L., Baron, J., Ramos, J., Gurcay, B., Fincher, K., Scott, S. E., Moore, D., Atanasov, P., Swift, S. A., Murray, T., Stone, E., & Tetlock, P. E. (2014). Psychological strategies for winning a geopolitical forecasting tournament. *Psychological Science, 25*(5), 1106-1115. https://doi.org/10.1177/0956797614524255

Mellers, B. A., Baker, J. D., Chen, E., Mandel, D. R., & Tetlock, P. E. (2017). How generalizable is good judgment? A multi-task, multi-benchmark study. *Judgment and Decision Making, 12*(4), 369-381.

Mellers, B. A., Tetlock, P. E., Baker, J. D., Friedman, J., & Zeckhauser, R. (2018). How much does predictive accuracy suffer when probability assessments are constrained? In H. Kunreuther, R. Meyer, & E. Michel-Kerjan (Eds.), *The future of risk management*. University of Pennsylvania Press.

Merkle, E. C., Steyvers, M., Mellers, B., & Tetlock, P. E. (2017). A neglected dimension of good forecasting judgment: The questions we choose also matter. *International Journal of Forecasting, 33*(4), 817-832. https://doi.org/10.1016/J.IJFORECAST.2017.04.002

Moore, D. A., Swift, S. A., Minster, A., Mellers, B., Ungar, L., Tetlock, P., Yang, H. H. J., & Tenney, E. R. (2017). Confidence calibration in a multiyear geopolitical forecasting competition. *Management Science, 63*(11), 3552-3565. https://doi.org/10.1287/MNSC.2016.2525

Satopää, V., Salikhov, M., Tetlock, P., & Mellers, B. (2022). Decomposing the effects of crowd-wisdom aggregators: The Bias-Information-Noise (BIN) model. *International Journal of Forecasting,* Advance online publication.

Satopää, V. A., Baron, J., Foster, D. P., Mellers, B. A., Tetlock, P. E., & Ungar, L. H. (2014). Combining multiple probability predictions using a simple logit model. *International Journal of Forecasting, 30*(2), 344-356. https://doi.org/10.1016/J.IJFORECAST.2013.09.009

Satopää, V. A., Jensen, S. T., Mellers, B. A., Tetlock, P. E., & Ungar, L. H. (2014). Probability aggregation in time-series: Dynamic hierarchical modeling of sparse expert beliefs. *Annals of Applied Statistics, 8*(2), 1256-1280. https://doi.org/10.1214/14-AOAS739

Satopää, V. A., Salikhov, M., Tetlock, P. E., & Mellers, B. (2021). Bias, information, noise: The BIN model of forecasting. *Management Science, 67*(12), 7599-7618. https://doi.org/10.1287/MNSC.2020.3882

Schwartz, H. A., Rouhizadeh, M., Bishop, M., Tetlock, P., Mellers, B., & Ungar, L. H. (2017). *Assessing objective recommendation quality through political forecasting* [Paper presentation]. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen.

Taleb, N. N. (2007). *The black swan: The impact of the highly improbable*. Random House.

Taleb, N. N., Bar-Yam, Y., & Cirillo, P. (2022). On single point forecasts for fat-tailed variables. *International Journal of Forecasting*, *38*(2), 413-422. https://doi.org/10.1016/J.IJFORECAST.2020.08.008

Taleb, N. N., & Tetlock, P. E. (2013). On the difference between binary prediction and true exposure with implications for forecasting tournaments and decision making research. *SSRN Working Paper*, Article 2284964. https://doi.org/10.2139/SSRN.2284964

Tetlock, P. E. (2017). *Expert political judgment: How good is it? How can we know?* Princeton University Press.

Tetlock, P. E., & Gardner, D. (2016). *Superforecasting: The art and science of prediction*. Random House.

Tetlock, P. E., Mellers, B. A., Rohrbaugh, N., & Chen, E. (2014). Forecasting tournaments: Tools for increasing transparency and improving the quality of debate. *Current Directions in Psychological Science, 23*(4), 290-295. https://doi.org/10.1177/0963721414534257

Tetlock, P. E., Mellers, B. A., & Scoblic, J. P. (2017). Bringing probability judgments into policy debates via forecasting tournaments. *Science, 355*(6324), 481-483. https://doi.org/10.1126/SCIENCE.AAL3147